ezCoref: Towards Unifying Annotation Guidelines for Coreference Resolution

Ankita Gupta Marzena Karpinska Wenlong Zhao Kalpesh Krishna
Jack Merullo Luke Yeh Mohit Iyyer Brendan O’Connor

University of Massachusetts Amherst Brown University Google
{ankitagupta, mkarpinska, wenlongzhao, kalpesh, miyier, brenocon}@cs.umass.edu
john_merullo@brown.edu, lukeyeh@google.com

Abstract

Large-scale, high-quality corpora are critical for advancing research in coreference resolution. However, existing datasets vary in their definition of coreferences and have been curated via complex and lengthy guidelines that are curated for linguistic experts. These concerns have sparked a growing interest among researchers to curate a unified set of guidelines suitable for annotators with various backgrounds. In this work, we develop a crowdsourcing-friendly coreference annotation methodology, ezCoref, consisting of an annotation tool and an interactive tutorial. We use ezCoref to re-annotate 240 passages from seven existing English coreference datasets (spanning fiction, news, and multiple other domains) while teaching annotators coreference cases typically handled by linguists only by training them how to deal with head-sharing noun phrases, premodifiers, and generic pronouns (appositives) in existing datasets. We propose the research community should revisit these phenomena when curating future unified annotation guidelines.

1 Introduction

Coreference resolution is the task of identifying and clustering together all textual expressions (mentions) that refer to the same discourse entity in a given document. Impressively, many have been made in developing coreference systems (Lee et al., 2017; Moosavi and Strube, 2018; Joshi et al., 2020), enabled by datasets annotated by experts (Hovy et al., 2006; Bamman et al., 2020; Uryupina et al., 2019) and crowdsourcing (Chamberlain et al., 2016). However, these datasets vary widely in their definitions of coreference (expressed via annotation guidelines), resulting in inconsistent annotations both within and across domains and languages. For instance, as shown in Figure 1, while ARR (Uryupina et al., 2019) treats generic pronouns as non-referential, OntoNotes (Hovy et al., 2006) chooses not to mark them at all.

It is thus unclear which guidelines one should employ when collecting coreference annotations in a new domain or language. Traditionally, existing guidelines have leaned towards lengthy explanations of complex linguistic concepts, such as those in the OntoNotes guidelines (Weischedel et al., 2012), which detail what should and should not be coreferent (e.g., how to deal with head-sharing noun phrases, premodifiers, and generic mentions). As a result, coreference datasets have traditionally been annotated by linguists (experts) already familiar with such concepts, which makes the process expensive and time-consuming. Crowdsourcing-friendly, experimental, and less intense guidelines are thus needed.

Figure 1: We visualize a common sentence from the news domain annotated by two expert-curated datasets, OntoNotes (Hovy et al., 2006) and ARR (Uryupina et al., 2019), along with the crowd annotations collected via our ezCoref platform. OntoNotes does not mark generic pronouns. ARR does not consider them as coreferent and annotates them using a special relation “undef-reference” (markables with vague interpretations). On the contrary, our crowdworkers assign all mentions of the generic pronoun “you” to the same coreference chain. The situation is also similar for the generic “we.”
sourced coreference data collection has the potential to be significantly cheaper and faster; however, teaching an exhaustive set of linguistic guidelines to non-expert crowd workers remains a formidable challenge. As a result, there has been a growing interest among researchers in curating a unified set of guidelines (Poensio et al., 2021) suitable for annotators with various backgrounds.

More recently, games-with-a-purpose (GWAPs) (von Ahn, 2006; Poensio et al., 2013) were proposed to aid crowdsourcing of large coreference datasets (e.g., Chamberlain et al., 2016; Yu et al., 2022). While GWAPs make it enjoyable for crowdworkers to learn complex guidelines and perform annotations using them (Madge et al., 2019b), they also require significant effort to attract and maintain workers. For instance, Phrase Detectives Corpus 1.0 was collected over a span of six years (Chamberlain et al., 2016; Poensio et al., 2013; Yu et al., 2022), which motivates us to instead study coreference collection on more efficient payment-based platforms.

Specifically, our work investigates the quality of crowdsourced coreference annotations when annotators are taught only simple coreference cases that are treated uniformly across existing datasets (e.g., pronouns). By providing only these simple cases, we are able to teach the annotators the concept of coreference, while allowing them to freely interpret cases treated differently across the existing datasets. This setup allows us to identify cases where our annotators unanimously agree with each other but disagree with the expert, thus suggesting cases that should be revisited by the research community when curating future guidelines.

Our main contributions are:

- We develop a crowdsourcing-friendly coreference annotation methodology—ezCoref—which includes an intuitive, open-sourced annotation tool supported by a short crowd-oriented interactive tutorial.\(^2\)
- We use ezCoref to re-annotate 240 passages from seven existing English coreference datasets on Amazon Mechanical Turk (AMT), and conduct a comparative analysis of crowd and expert annotations. We find that high-quality annotations are already achievable from non-experts without extensive training (>90% B3 (Bagga and Baldwin, 1998a) agreement between crowd and experts).
- We further qualitatively analyze remaining disagreements among crowd and expert annotations and identify linguistic cases that crowd unanimously marks as coreferent but lack unified treatment in existing datasets (e.g., generic pronouns as shown in Figure 1). Additionally, analyzing inter-annotator agreement among the crowd reveals that crowd exhibits higher agreement when annotating familiar texts (e.g., childhood stories or fiction) compared to texts rich in cataphoras or those requiring world knowledge. Finally, our qualitative analysis also provides an empirical evidence to support previous findings in literary studies (Szakolczai’s (2016) analysis of Bleak House) and psychology (Orvell et al.’s (2020) claims about generic “you”).

2 Related Work

Existing coreference datasets: Table 1 provides an overview of seven prominent coreference datasets, which differ widely in their annotator population, mention detection, and coreference guidelines.\(^3\) Many datasets are annotated by experts heavily trained in linguistic standards, including ARRAU (Uryupina et al., 2019), LitBank (Bamman et al., 2020), GUM (Zeldes, 2017), and OntoNotes (Hovy et al., 2006). Due to its scale and quality, OntoNotes is likely the most widely used for NLP coreference research, including in two CoNLL shared tasks (Pradhan et al., 2011, 2012). QuizBowl (Guha et al., 2015) has been annotated by domain (but not linguistic) experts. Few coreference datasets exist which are annotated by non-experts, including those created by part-time non-native English speakers (PreCo; Chen et al., 2018), and gamified crowdsourcing without financial compensation (Phrase Detectives; Chamberlain et al., 2016; Yu et al., 2022).

Coreference annotation tools: Several coreference annotation tools have been developed (See Table A3 in Appendix for more details). However, these are difficult to port to a crowdsourced workflow, as they require users to install software on their local machine (Widlöcher and Mathet, 2012; Landragn et al., 2012; Kopeć, 2014; Mueller and Strube, 2001; Reiter, 2018), or have complicated

\(^2\)Our tutorial received overwhelmingly positive feedback. One annotator commented that it was “absolutely beautiful, intuitive, and helpful. Legitimately the best one I’ve ever seen in my 2 years on AMT! Awesome job.” (Table A4 in Appendix)

\(^3\)Many others exist too; for example, see Jonathan Kummerfeld’s spreadsheet list (accessed Jan. 2022).
Table 1: Summary of seven datasets analyzed in this work, which differ in domain, size, annotator qualifications, mention detection procedures, types of mentions, and types of links considered as coreferences between these mentions. *Allows other types of mention only when this mention is an answer to a question. **We interpret manual identification based on illustrations presented in the original publication (Chen et al., 2018). ***Inaccessible, see Footnote 8.

UI design with multiple drag and drop actions and/or multiple windows (Stenetorp et al., 2012; Widlocher and Mathet, 2012; Landragin et al., 2012; Yimam et al., 2013; Girardi et al., 2014; Kopeć, 2014; Mueller and Strube, 2001; Oberle, 2018). Closest to ezCoref is CoRefi (Bornstein et al., 2020), a web-based coreference annotation tool that can be embedded into crowdsourcing websites. Subjectively, we found its user interface difficult to use (e.g., users have to memorize multiple key combinations). It also does not allow for nested spans, reducing its usability.

Crowdsourcing linguistic annotations: Several efforts have been made to crowdsource linguistic annotations (Snow et al., 2008; Callison-Burch, 2009; Howe, 2008; Lawson et al., 2010), including on payment-based microtasks via platforms like AMT and GWAPs (von Ahn, 2006). Many GWAPs (Poesio et al., 2013; Kicikoglu et al., 2019; Madge et al., 2019b; Miller et al., 2019), where annotators start with simpler tasks and gradually move towards more complex problems, but this requires subjective judgments of task difficulty. In contrast to the payment-based microtask setting studied in this work, GWAPs are not open-sourced, need significant development, take longer to collect data, and require continuous efforts to maintain visibility (Poesio et al., 2013).

3 ezCoref: A Crowdsourced Coreference Annotation Platform

The ezCoref user experience consists of (1) a step-by-step interactive tutorial and (2) an annotation interface, which are part of a pipeline including automatic mention detection and AMT Integration.

Annotation structure: Two annotation approaches are prominent in the literature: (1) a local pairwise approach, annotators are shown a pair of mentions and asked whether they refer to the same entity (Hladká et al., 2009; Chamberlain et al., 2016; Li et al., 2020; Ravenscroft et al., 2021), which is time-consuming; or (2) a cluster-based approach (Reiter, 2018; Oberle, 2018; Bornstein et al., 2020), in which annotators group all mentions of the same entity into a single cluster. In ezCoref we use the latter approach, which can be faster but requires the UI to support more complex actions for creating and editing cluster structures.
Table 2: Simple coreference cases explained in tutorial.

<table>
<thead>
<tr>
<th>Example</th>
<th>Phenomena Taught</th>
</tr>
</thead>
<tbody>
<tr>
<td>[John] doesn’t like [Fred], but [he] still invited [him] to [the party].</td>
<td>(1) personal pronouns</td>
</tr>
<tr>
<td>[This dog] likes to play [catch].</td>
<td>(1) possessive pronouns</td>
</tr>
<tr>
<td>[Director [Mackenzie]] spent [last two years] working on a [&quot;Young Adam&quot;].</td>
<td>(1) nested spans</td>
</tr>
<tr>
<td>[The office] wasn’t exactly small either.</td>
<td>(1) non-person entities (place)</td>
</tr>
</tbody>
</table>

User interface: We spent two years iteratively designing, implementing, and user testing the interface to make it as simple and crowdsourcing-friendly as possible (Figure 2). Marked mentions are surrounded by color-coded frames with entity IDs. The currently selected mention (“the book”), is highlighted with a flashing yellow cursor-like box. The core annotation action is to select other mentions that corefer with the current mention, and then advance to a later unassigned mention; annotators can also re-assign a previously annotated mention to another cluster. Advanced users can exclusively use keyboard shortcuts, undo and redo actions were added to allow error correction. Finally, ezCoref provides a side panel showing mentions of the entity currently being annotated to spot mentions assigned to the wrong cluster.

Coreference tutorial: To teach crowdworkers the basic definition of coreference and familiarize them with the interface, we develop a tutorial (aimed to take ~ 20 minutes) that introduces them to the mechanics of the annotation tool, and then trains them on simple cases of coreferences. These cases (e.g., personal/possessive pronouns or determinative phrases that corefer with their antecedents as shown in Table 2) are annotated similarly across all existing datasets and are unlikely to be disputed. The tutorial concludes with a quality control example to exclude poor quality annotators. These training examples, feedback, and annotation guidelines can be easily customized using a simple JSON schema.

Annotation workflow: The annotators are presented with one passage (or “document”) at a time (Figure 2), and all mentions have to be annotated before proceeding to the next passage. There is no limitation to the length or language of the passage.

In this work, we divide an initial document into a sequence of shorter passages of complete sentences, on average 175 tokens, as shorter passages minimize the need to scroll, reducing annotator effort. While this obviously cannot capture longer distance coreference, a large portion of important coreference phenomena is local: within the OntoNotes written genres, for pronounial mentions, the closest antecedent is contained within the current or previous two sentences more than 95% of the time.

Automatic mention detection: As a first step to collect coreference annotations, we must identify mentions in the documents from each of the seven existing datasets; this process is done in a diverse array of ways (from manually to automatic) in prior work as shown in Table 1. We decided to automatically identify mentions to give all crowdworkers an identical set of mentions, which simplifies the annotation task and also allows us to easily compare and study their coreference annotations via inter-annotator agreement. Specifically, we implement a simple algorithm that yields a high average recall over all seven datasets.

Our algorithm considers all noun phrases (including proper nouns, common nouns, and pronouns) as markables, extracting them using the Stanza dependency parser (version 1.3.0; Qi et al., 2020). We allow for nested mentions and proper noun premodifiers (e.g., [U.S.] in “U.S. policy”). We include all conjuncts with the entire coordinated noun phrase ([Mark], [Mary], as well as [Mark and Mary], all considered mentions); details in Appendix A.3.

4 Using ezCoref to Re-annotate Existing Coreference Datasets

We deploy ezCoref on the AMT crowdsourcing platform to re-annotate 240 passages from seven existing datasets, covering seven unique domains. In total, we collect annotations for 12,200 mentions and 42,108 tokens. We compare our workers’ annotations with one passage (or “document”) at a time (Figure 2), and all mentions have to be annotated before proceeding to the next passage. There is no limitation to the length or language of the passage.

6We leave this for future work—for example, more sophisticated user interfaces to support longer documents, or merging coreference chains between short passages. As documents get progressively longer, such as book chapters or books, the task takes on aspects of cross-document coreference and entity linking (e.g. Bagga and Baldwin, 1998b; FitzGerald et al., 2021; Logan IV et al., 2021).

7We acknowledge that any algorithm can be used as long as its recall across all datasets is high, and ours is only one such algorithm. However, we do not conduct an ablation study to compare crowd annotations for mentions obtained from these potential algorithms as it would be prohibitively expensive. Furthermore, while advanced mention detection methods can improve annotation quality, our goal is not to collect the highest-quality coreference dataset, but to study annotator behavior when a common set of mentions is provided.
notations both quantitatively and qualitatively to each other and to existing expert annotations.

Datasets: We collect coreference annotations for the seven existing datasets described in Table 1: OntoNotes (Hovy et al., 2006), LitBank (Bamman et al., 2020), PreCo (Chen et al., 2018), ARRRAU (Uryupina et al., 2019), GUM (Zeldes, 2017), Phrase Detectives (Chamberlain et al., 2016), and QuizBowl (Guha et al., 2015). The sample covers seven domains: news, opinionated magazines, weblogs, fiction, biographies, Wikipedia articles, and trivia questions from Quiz Bowl. For each dataset with multiple domains, we manually select a broad range of domain(s) for re-annotation. From each domain in each dataset, we then select documents and divide them into shorter passages (on average 175 tokens each), creating 20 such passages per dataset. For datasets with multiple domains, we choose 20 such passages per domain (see Appendix A.1 for detail). Overall, we collect annotations for 240 passages with 5 annotations per passage to measure inter-annotator agreement.

Procedure: We first launch an annotation tutorial and recruit the annotators on the AMT platform. At the end of the tutorial, each annotator is asked to annotate a short passage (around 150 words). Only annotators with a B3 score (Bagga and Baldwin, 1998a) of 0.90 or higher are then invited to participate in the annotation task.

Training Annotators with Simplified Guidelines using ezCoref: As the goal of our study is to understand what crowdworkers perceive as coreference, we train our annotators with simple guidelines. We carefully draft our training examples to include only cases which are considered as coreference by all the existing datasets. The objective is to teach crowdworkers the broad definition of coreference while leaving space for different interpretations of ambiguous cases or those resolved differently across the existing datasets. Note that a comparable experiment with more complex guidelines is infeasible since it is unclear which guidelines to choose, and also providing complex linguistic guidelines to crowdworkers remains an open challenge. Overall, ezCoref is aimed to minimize both researcher and annotator effort for new coreference data collection, compared to prior work (Figure 3).

Worker details: Overall, 73 annotators (including 44 males, 20 females, and one non-binary person) completed the tutorial task, which took 19.4 minutes on average (sd=11.2 minutes). They were aged between 21 and 69 years (mean=38.9, sd=11.3) and identified themselves as native English speakers. Most of the annotators had at least a college degree (47 vs 18). 89.0% of annotators, who did the tutorial, received a B3 score of 0.90 or higher for the final screening example, and were invited to the annotation task. 50.7% of the invited annotators returned to participate in the main annotation task, and 29.2% of them annotated five or more passages. Annotation of one passage took, on average, 4.15 minutes, a rate of 2530 tokens per hour. The total cost of the tutorial was $460.70 ($4.50 per tutorial). We paid $1 per passage for the main annotation task, resulting in a total cost of $1440.11

5 Analysis
In this section, we perform quantitative and qualitative analyses of our crowdsourced coreference annotations. First, we evaluate the performance

---

11The PreCo dataset is interestingly large but seems difficult to access. In November 2018 and October 2021 we filled out the data request form at the URL provided by the paper, and attempted to contact the PreCo official email directly, but did not receive a response. To enable a precise research comparison, we scraped all documents from PreCo’s public demo in November 2018 (no longer available as of 2021); its statistics match their paper and our experiments use this version of the data. PreCo further suffers from data curation issues (Gebru et al., 2018; Jo and Gebru, 2020); it uses text from English reading comprehension tests collected from several websites, but the original document sources and copyright statuses are undocumented. When reading through PreCo documents, we found many domains including opinion, fiction, biographies, and news (Table A1 in Appendix); we use our manual categories for domain analysis.

10We allow only workers with a >= 99% approval rate and at least 10,000 approved tasks who are from the US, Canada, Australia, New Zealand, or the UK.

11We did not collect demographic data for the remaining eight individuals, from an earlier pilot experiment.

11All reported costs include 20% AMT fee.
of our mention detection algorithm, comparing it to gold mentions across seven datasets. Next, we measure the quality of our annotations and their agreement with other datasets. Finally, we discuss interesting qualitative results.

5.1 Mention Detector Evaluation

Datasets differ in the way they define their mention boundaries and thus the boundaries for the same mention may differ. To fairly compare our mentions with the gold standards, we employ a headword-based comparison. We find the head of the given phrase by identifying, in the dependency tree, the most-shared ancestor of all tokens within the given mention. Two mentions are considered same if their respective headwords match.

Table 3 compares our mention detector to the gold mentions in existing datasets. Our method obtains high recall across most datasets (>0.90), which shows that most of the mentions annotated in existing datasets are correctly identified and allows a direct comparison of crowd annotations with expert annotations. It has the lowest recall with ARRAU (0.84) and PreCo (0.88), which is to be expected as ARRAU marks all referring premodifiers (identified manually) and PreCo allows common noun modifiers, while we identify only the premodifiers which are proper nouns.12

For most datasets, the precision is >0.80, suggesting that the algorithm identifies most of the relevant mentions. We observe a substantially lower score for OntoNotes, LitBank, and QuizBowl as these datasets restrict their mention types to limited entities (refer to Table 1). However, this does not limit our analysis. In fact, an algorithm with high precision on LitBank or OntoNotes would miss a huge percentage of relevant mentions and entities on other datasets (constraining our analysis) and when annotating new texts and domains. Furthermore, our algorithm identifies more mentions than in the original datasets, which in the best case allows us to discover new entities and, in the worst case, may result in more singletons. Finally, the mention density (number of mentions per token) from our detector remains roughly consistent across all datasets when using our method, allowing us to fairly compare statistics (e.g., agreement rates) across datasets.

Table 3: Comparison of mentions identified by our mention detection algorithm with the gold mentions annotated in the respective datasets. We use head-word based comparison to compare mentions of different lengths. Our method obtains high recall across most datasets and the mention-density using our mention-detector remains roughly consistent across datasets, allowing us to do fair analysis (e.g., agreement) across datasets.

5.2 Agreement with Existing Datasets

How well do annotations from ezCoref agree with annotations from existing datasets?

Aggregating annotations: To compare crowdsourced annotations with gold annotations, we first require an aggregation method that can combine annotations from multiple crowdworkers to infer coreference clusters. We use a simple aggregation method that determines whether a pair of mentions is coreferent by counting the number of annotators who marked the two mentions in the same cluster.13 Two mentions are considered as coreferent when the number of annotators linking them together is greater than a threshold (τ). After inferring these pairs of mentions, we construct an undirected graph where nodes are mentions and edges represent coreference links. Finally, we find connected components in the graph to obtain coreference clusters.14 We compare aggregated annotations from ezCoref with gold annotations across the seven datasets using B3 scores (precision, recall, and F1),15 as illustrated in Figure 4.

High agreement with OntoNotes, GUM, LitBank, ARRAU: Our annotators achieve the high-

---

12We made this decision as identifying automatically all premodifiers would result in many singletons and lead to more arduous annotation effort.

13Future data collection efforts interested in creating large resources can utilize more advanced aggregation methods (Poesio et al., 2019).

14This method resolves to majority voting-based aggregation when the τ is set so that more than half of annotators should agree. For τ = N, this method is very conservative, adding a link between two mentions only when all annotators agree unanimously. Conversely, for τ = 1, only a single vote is required to add a link between two mentions.

15For a mention in a given document, B3 recall is the fraction of mentions that are correctly predicted by the system as coreferent with it out of all mentions that are actually coreferent with it. B3 precision is the fraction of mentions that are correctly predicted by the system as coreferent with it out of all system-predicted mentions.
est precision with OntoNotes (Figure 4), suggesting that most of the entities identified by crowd-workers are correct for this dataset. In terms of F1 scores, the datasets which are closest to crowd annotations are GUM, LitBank, and ARRAU, all of which are annotated by experts. This result shows that high-quality annotations can be obtained from non-experts using ezCoref without extensive training. We further conducted a qualitative analysis of high agreement cases for each dataset. Overall, we observe that non-experts agree with experts on chains containing pronouns and named entities. However, non-experts also mark noun phrases in appositive constructions as coreferent, consistent with GUM guidelines. Finally, non-experts also assign generic mentions to the same coreference chain, consistent with their treatment by GUM and ARRAU, and leads to higher agreement with these datasets.

Low precision with Phrase Detectives and PreCo, low recall with Quiz Bowl: We observe that Phrase Detectives has a very low precision compared to all other datasets, implying that crowd-workers add more links compared to gold annotations. Our qualitative analysis reveals that PD annotators miss some valid links, splitting entities which are correctly linked together by our annotators (see Table 4). Another dataset with lower precision is PreCo, which also contains many missing links. In general, we observe more actual mistakes in PreCo and PD than in the other datasets, which is not surprising as they were not annotated by experts. This result is further validated by our agreement analysis of the fiction domain (Table 5), in which ezCoref annotations agree far more closely with expert annotations (GUM, LitBank) than PreCo and PD. Finally, Quiz Bowl has by far the lowest recall with ezCoref annotations, which is expected given the difficulty with cataphora and factual knowledge (examples (c) and (e) in Table 6).

Table 4: Cases of split entities (missing links) in annotations provided with Phrase Detectives and PreCo. Instead, our crowd annotators mark all mentions as referring to the same entity in each of these examples.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dataset</th>
<th>B3 Precision</th>
<th>B3 Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction</td>
<td>GUM</td>
<td>0.982</td>
<td>0.921</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>LitBank</td>
<td>0.959</td>
<td>0.927</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td>PreCo</td>
<td>0.805</td>
<td>0.963</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>Phrase Detectives</td>
<td>0.784</td>
<td>0.775</td>
<td>0.780</td>
</tr>
</tbody>
</table>

Table 5: Agreement with existing datasets for fiction.

Varying the aggregation threshold $\tau$: What is the effect of varying the aggregation threshold ($\tau$) on precision and recall with gold annotations? Figure 5 shows that the Quiz Bowl dataset has the highest drop in recall (36% absolute drop) when increasing $\tau$ from 1 to 5. This indicates that the number of unanimous clusters ($\tau = 5$) is considerably lower than the total number of clusters found individually by all annotators ($\tau = 1$); as such, our annotators heavily disagree about gold clusters in the Quiz Bowl dataset. We observe a similar trend in OntoNotes (26% drop in recall), whereas Phrase Detectives has the lowest drop in recall (0.07) with the increase in the number of annotators, which is expected since Phrase Detectives is crowdsourced.

5.3 What domains are most suitable for crowdsourcing coreference?

We use the B3 metric (Bagga and Baldwin, 1998a) to compute IAA for each domain, excluding singletons (see Table 7). We obtain the highest agreement on fiction (72.6%) and biographies (72.4%). This is because both domains contain a high frequency of pronouns (see examples a and b).

---

16That said, both PreCo and PD were additionally validated by multiple non-expert annotators.

17We analyze variations in recall which is more interpretable than precision, since the denominator is fixed in recall when varying number of annotators.

18Krippendorff’s alpha/kappa are other possible measures for IAA. However, prior work (Paun et al., 2022) has raised concerns over using Krippendorff’s alpha/kappa for anaphora resolution. Instead, we found B3 intuitive to understand as a measure of agreement among annotators at the mention level, i.e. fraction of mentions two annotators agree should be coreferent with a given mention.

19IAA including singletons is much higher (Appendix A4).
We find that **ezCoref** we look at each link that was annotated by our (e.g., Little Red Riding Hood) that are likely familiar to our annotators, which may have made them easier to annotate. Annotators have the least agreement on Quiz Bowl coreference (59.73%), as this dataset is rich in challenging cataphoras (example **c** in Table 6) and often require world knowledge about books, characters, and authors to identify coreferences (example **e** in Table 6).  

### 5.4 Qualitative analysis

To better understand the differences in annotation quality, we conduct a manual analysis of all 240 passages, comparing our ezCoref annotations to gold annotations from each dataset. Specifically, we look at each link that was annotated by our workers but not in the gold data, or vice versa. For each link, we determine whether crowd or the gold annotations contained a mistake, or whether the discrepancy is reasonable under specific guidelines. We find that ezCoref annotations contain fewer mistakes than non-expert annotated datasets (PreCo and PD), almost twice as many mistakes as those of expert datasets (OntoNotes and GUM), and seven times as many mistakes as those in the esoteric Quiz Bowl dataset (Appendix Table A2).

### Disagreements and deviations from expert guidelines:

As in Poesio and Artstein (2005), we identify cases of genuine ambiguity, where a mention can refer to two different antecedents. This first row of Table 8 shows an example from Dickens’ *Bleak House*, where the pronoun “it” could reasonably refer to either the “fog” or the “river.” Our annotators have high disagreement on this link, which is understandable given the literary analysis of Szakolczai (2016) who interprets the ambiguity of this pronoun as Dickens’ way to show indeterminacy attributed to elements in the scene.  

We observe that generic mentions, especially generic pronouns, are almost always annotated as coreferring by crowd, while existing datasets lack consensus (Table 1). Table 8 (second row) shows an example where annotators unanimously connected all instances of generic “you.” This observation is in line with Orvell et al.’s (2020) study which explains that by using the same linguistic form (“you”), one invites readers (annotators) to consider how the situation refers to them. Finally, while datasets tend to treat copulae and appositive constructions identically and annotate them...
Ambiguity

[Fog] everywhere. [Fog] up [the river], where [it] flows among green

[Generic]

Table 8: Examples of genuine ambiguity and generic “you” observed in our data.

in a similar way, our annotators intuitively annotate them differently. While crowdworkers almost always mark noun phrases in appositive constructions as coreferent, the noun phrases in copulas are linked by majority vote only in ∼ 35% of cases.

6 Conclusion

Existing coreference datasets vary in their definition of coreferences and have been collected via complex guidelines. In this work, we investigate the quality of annotations when crowdworkers are taught only few coreference cases that are treated similarly across existing datasets. We develop a crowdsourcing-friendly coreference annotation methodology, ezCoref and use it to re-annotate 240 passages from seven existing English coreference datasets. We observe reasonable quality annotations were already achievable even without extensive training. On analyzing the remaining disagreements, we identify linguistic cases that crowd unanimously agree upon but lack unified treatments in existing datasets, suggesting cases the researchers should revisit when curating future unified annotation guidelines.

7 Limitations

We list some of the limitations of our study which researchers and practitioners would hopefully benefit from when interpreting our analysis. Firstly, our analysis is only applicable to the English language and how native English speakers understand coreferences. In this work, we have taken a step towards building a framework to facilitate the comparison of the crowd and expert annotations, and the variations observed in non-native speakers should be explored in future studies. Secondly, as a result of resource constraints, we limited ourselves to one set of guidelines and compared crowd annotations under these guidelines with expert annotations. Understanding the effects of various guidelines on annotator behavior is left for future research. Thirdly, even the best automatic mention detection algorithm could have errors, especially when tested out-of-domain. Despite this limitation, we decided to use an automatic method as it allows us to study annotators’ behavior when a “common set of mentions” is provided. Some of the proposed solutions to address this issue are to directly crowdsource mentions or verify the automatically identified mentions via crowdsourcing (Madge et al., 2019b), which can be utilized for future collection of high-quality corpora. Finally, we also acknowledge that the tool cannot handle split-antecedents or separate tags for different relations, which we leave for future work. As a result, our approach focuses on cases of identity coreferences. However, we believe that identity coreference supported by our tool has value as an NLP tool (e.g., studying characters in narratives (Bamman et al., 2013)), allowing the collection of more in-domain annotations, necessary to advance such practical applications.

8 Ethics Statement

The data collection protocol was approved by the coauthors’ institutional review board. All annotators were presented with a consent form (mentioned below) prior to the annotation. They were also informed that only satisfactory performance on the screening example will allow them to take part in the annotation task. All data collected during the tutorial and annotations (including annotators’ feedback and demographics) will be released anonymized. We also ensure that the annotators receive at least $13.50 per hour. Since base compensation is per unit of work, not by time (the standard practice on Amazon Mechanical Turk), we add bonuses for workers whose speed caused them to fall below that hourly rate.

Consent Before participating in our study, we requested every annotator to provide their consent. The annotators were informed about the purpose of this research study, any risks associated with it, and the qualifications necessary to participate. The consent form also elaborated on task details describing what they will be asked to do and how long it will take. The participants were informed that they could choose as many documents as they would like to annotate (by accepting new Human Intelligence Tasks at AMT) subject to availability, and they may drop out at any time. Annotators were informed that they would be compensated in the standard manner through the Amazon Mechanical Turk crowdsourcing platform, with the amount specified in the Amazon Mechanical Turk interface. As part of this study, we also collected demographic information, including their age, gender,
native language, education level, and proficiency in the English language. We ensured our annotators that the collected personal information would remain confidential in the consent form.

Acknowledgements

We are very grateful to the crowd annotators on AMT for participating in our annotation tasks and providing positive reviews. We are grateful to Abe Handler, Aditya Jain, Anna Rogers, Julian Richardson, Kavya Jeganathan, Neha Kennard, Nishant Yadav, Timothy O’Gorman, and the UMass NLP group for several useful discussions during the course of the project. We also thank Massimo Poesio for sharing the GNOME portion of AR-RAU dataset. This material is based upon work supported by National Science Foundation awards 1925548, 1814955, and 1845576, and a Google PhD Fellowship awarded to KK.

References


A Appendix

A.1 Details of our crowdsourced data

Table A1 mentions all datasets that we re-annotate in this work with their breakdown based on domains, number of documents, passages, tokens and mentions annotated.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>#Docs</th>
<th>#Passages</th>
<th>#Tokens</th>
<th>#Mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>OntoNotes</td>
<td>News</td>
<td>6</td>
<td>30</td>
<td>4923</td>
<td>1365</td>
</tr>
<tr>
<td></td>
<td>Weblogs</td>
<td>5</td>
<td>20</td>
<td>3452</td>
<td>1001</td>
</tr>
<tr>
<td></td>
<td>Opinion</td>
<td>12</td>
<td>20</td>
<td>3861</td>
<td>1157</td>
</tr>
<tr>
<td>LitBank</td>
<td>Fiction</td>
<td>4</td>
<td>30</td>
<td>5455</td>
<td>1494</td>
</tr>
<tr>
<td>QuizBowl</td>
<td>Quizzes</td>
<td>20</td>
<td>20</td>
<td>3304</td>
<td>1083</td>
</tr>
<tr>
<td>ARRAU</td>
<td>News</td>
<td>3</td>
<td>20</td>
<td>3336</td>
<td>885</td>
</tr>
<tr>
<td>GUM</td>
<td>Biographies</td>
<td>4</td>
<td>20</td>
<td>3422</td>
<td>1119</td>
</tr>
<tr>
<td></td>
<td>Fiction</td>
<td>4</td>
<td>20</td>
<td>3299</td>
<td>1008</td>
</tr>
<tr>
<td>Phrase</td>
<td>Wikipedia</td>
<td>7</td>
<td>20</td>
<td>3509</td>
<td>1003</td>
</tr>
<tr>
<td>Detectives</td>
<td>Fiction</td>
<td>4</td>
<td>20</td>
<td>4007</td>
<td>1063</td>
</tr>
<tr>
<td>Opinion</td>
<td>9</td>
<td>18</td>
<td>1692</td>
<td>495</td>
<td></td>
</tr>
<tr>
<td>PreCo</td>
<td>News</td>
<td>4</td>
<td>8</td>
<td>1318</td>
<td>369</td>
</tr>
<tr>
<td></td>
<td>Fiction</td>
<td>2</td>
<td>2</td>
<td>378</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Biographies</td>
<td>1</td>
<td>1</td>
<td>152</td>
<td>53</td>
</tr>
</tbody>
</table>

Total All 83 240 42108 12200

Table A1: All datasets analyzed in this work with their breakdown based on domains, number of documents, passages, tokens and mentions annotated.

A.2 Manual Qualitative Analysis

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Mistakes (our)</th>
<th>Mistakes (gold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD (silver)</td>
<td>22</td>
<td>76</td>
</tr>
<tr>
<td>OntoNotes</td>
<td>81</td>
<td>49</td>
</tr>
<tr>
<td>PreCo</td>
<td>12</td>
<td>33</td>
</tr>
<tr>
<td>GUM</td>
<td>48</td>
<td>25</td>
</tr>
<tr>
<td>ARRAU</td>
<td>33</td>
<td>16</td>
</tr>
<tr>
<td>LitBank</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>QuizBowl</td>
<td>67</td>
<td>10</td>
</tr>
</tbody>
</table>

Table A2: Number of mistakes in our crowd annotations vs. gold datasets, obtained through a manual analysis.

A.3 Detailed Mention Detection Algorithm

- We identify all noun phrases using the Stanza dependency parser (Qi et al., 2020). For each word with a noun-related part-of-speech tag, we recursively traverse all of its children in the dependency graph until a dependency relation is found in a whitelist. The maximal span considered as a candidate mention thus covers all words related by relations in the whitelist.

- Possessive nominal modifiers are also considered as candidate mentions. For instance, in the sentence “Mary’s book is on the table,” we consider both “Mary” and “Mary’s book” as mentions.

- Modifiers that are proper nouns in a multi-word expression are considered as mentions. For instance, in “U.S. foreign policy,” the modifier “U.S.” is also considered as a mention.

- All conjuncts, including the headword and other words depending on it via the conjunct relation, are considered mentions in a co-ordinated noun phrase. For instance, in the sentence, “John, Bob, and Mary went to the party,” the detected mentions are “John,” “Bob,” “Mary,” and the coordinated noun phrase “John, Bob, and Mary.”

- Finally, we remove mentions if a larger mention with the same headword exists. We allow nested spans (e.g., [[my] hands]) but merge any intersecting spans into one large span (e.g., [western [Canadian] province] is merged into [western Canadian province]).

A.4 Inter-Annotator Agreement Among Our Annotators Across Domains

Figure 6 illustrates agreement among our annotators computed with B3 scores including singletons.

![Inter Annotator Agreement across different domains](image)

Figure 6: Inter Annotator Agreement across different domains. B3 scores with Singletons included.

A.5 Another illustrative example

An example of a single sentence annotated by two datasets, OntoNotes and ARRAU. These annotations differ widely from each other in kinds of mentions and links between mentions.

**OntoNotes:** [ Lloyd’s, once a pillar of the world insurance market ]e1, e2 is being shaken to its very foundation.

**ARRAU:** [ Lloyd’s, once a pillar of the world insurance market ]e3 e2 e51 e1, is being shaken to its very foundation e52.

---

22Pronouns, nouns, proper nouns, and numbers.

23The whitelist includes all multi-word expression relations (i.e., compound, flat, and fixed) and modifier relations (i.e., determiners, adjectival modifiers, numeric modifiers, nominal modifiers, and possessive nominal modifiers).
Table A3: A comparison of different coreference annotation tools. (* — ezCoref code will be open-sourced upon paper publication; Stenetorp et al. (2012) did not implement nested spans originally, but later added them with limited functionality. Yimam et al. (2013) have APIs for CrowdFlower integration, but suggest expert annotators.).

*Accessible at: https://prodi.gy/

<table>
<thead>
<tr>
<th>System</th>
<th>Annotate all clusters</th>
<th>Pre-identified Mentions</th>
<th>Open Source</th>
<th>Webapp</th>
<th>Coref only</th>
<th>Keyboard and Mouse</th>
<th>MTurk Tested</th>
<th>Non-expert Terminology</th>
<th>Nested Span Support</th>
<th>Interactive Tutorial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stenetorp et al. (2012)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Widlocher and Mathet (2012)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Landragin et al. (2012)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Yimam et al. (2013)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Poesio et al. (2013)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Girardi et al. (2014)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mueller and Straube (2001)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kopec (2014)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Guha et al. (2015)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Obele (2016)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Reiter (2018)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bornstein et al. (2020)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Prodigy*</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ezCoref (this work)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table A4: Some of the comments received from our annotators after completing the tutorial. We received overwhelmingly positive feedback; annotators sometimes also mentioned cases they found confusing.
Coreference Tutorial

Welcome!
This is a paid tutorial for the "Large-Scale Coreference Annotation Task."
In this tutorial you will learn how to annotate coreferences, that is, words and phrases that refer to the same people or things.
Upon completing the tutorial, you will get a completion code. You MUST enter this code in the textbox below and submit the HIT in order to receive the payment.
Depending on your performance, you might be invited to participate in our "Large-Scale Coreference Annotation Task."

Before proceeding to the tutorial, please fill in the following survey:

What is your gender?

________

What is your age?

________

What is your native language?

________

How is your English level?

☐ Beginner
☐ Intermediate
☐ Advanced (near native)
☐ Native speaker

What is your education level?

☐ Primary
☐ Secondary
☐ College
☐ Graduate School

Click this link to begin.
[OPTIONAL] We would love to hear your feedback about this tutorial.

Submit your code below:

Submit

Figure 7: Screenshot of tutorial task invitation on AMT with detailed instructions.
Coreference Tutorial Mode

Welcome to the coreference tutorial mode. Here you will learn how to use the interface efficiently to label text for coreferences.

What are coreferences?
A coreference is when two words or spans (sequence of words) refer to the same thing.

In the examples below, the following words are coreferences (they refer to the same "thing"):  
1) "John" and "He"
2) "Robert" and "He"
3) "Alice" and "Her"

John is cool. He is nice.
Robert loves Alice. He talks to her everyday.

Let's get started.

Figure 8: Tutorial Interface (Introductory prompt)

Select Spans (Task 1 of 10)

Step 1 of 2

Observe how the border around "Mary" is flashing. This means the span "Mary" is the current target. Click on the spans that refer to the target "Mary".

Shortcuts

<table>
<thead>
<tr>
<th>Function</th>
<th>Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Target</td>
<td>a</td>
</tr>
<tr>
<td>Next Target</td>
<td>d</td>
</tr>
<tr>
<td>Undo</td>
<td>Ctrl + Z</td>
</tr>
<tr>
<td>Redo</td>
<td>Ctrl + Y</td>
</tr>
</tbody>
</table>

Figure 9: Tutorial interface: A sample prompt teaching tool functionality.
Figure 10: Tutorial interface: A sample prompt teaching basic coreferences.

Figure 11: Tutorial interface: quality control example.
Coreference Annotation Task

Welcome to the coreference annotation task. In this task you will be asked to annotate a short paragraph for coreferences. If you need to review the tutorial, please follow this link.

What are coreferences?

A coreference is when two words or spans (sequence of words) refer to the same thing.

In the examples below, the following words are coreferences (they refer to the same “thing”):

(1) "John" and "He"
(2) "Robert" and "He"
(3) "Alice" and "Her"

Click this link to begin annotation.

[OPTIONAL] We would love to hear your feedback. Let us know if anything was unclear or particularly challenging.

Submit your code below:

Submit

Figure 12: Annotation task invite on AMT with detailed instructions